M2 internship - 2021

Emotion Speech Signal Analysis for Geriatric Medicine

Mar. 2021 - Aug. 2021 (5-6 months)

Supervisor(s)	:	Dominique Fourer (with H. Maaref)
Team / Laboratory	:	SIAM / IBISC (EA 4526) - Univ. Évry/Paris-Sacay
Collaborators	:	LaBRI, IMS (Univ. Bordeaux)
Contact	:	dominique.fourer@univ-evry.fr

Abstract : "Humanitude" is a healing technique generalized by Gineste and Marescotti [5, 4] which claims to provide optimal communication skills in elder care facilities. This approach based on affective communication between health professionals and elderly patients has succeeded to improve the cognitive capabilities of fragile people hospitalized in the context of health care in EHPAD. This methodology based on affective speech is currently investigated in the MSH-HUMAVOX project which aims to objectively characterize and better understand why this approach can significantly improve the life quality and reduce behavioral disorders associated with senile state. Hence, this work focuses on the analysis of audio speech signal which showed its capability to convey relevant information about emotion and/or socio-cultural codes independently from the semantic content [8, 3]. The goal of this internship is to propose a complete analysis-synthesis system allowing to recognize emotion from audio speech recordings and to synthesize speech signals with a target emotion state.

keywords : speech signal processing, emotion analysis/synthesis, deep learning

Goals

— Bibliographical study for identifying the best state-of-the-art methods for emotion speech recognition and synthesis.
— Implementation of a new proposed technique for audio speech analysis-synthesis
— Analysis and interpretation of the emotion-relevant acoustic signal features

Methodology

The starting point of this research is our previous works based on the prosody analysis of social attitudes which showed the relevance of several acoustic parameters such as the fundamental frequency (F_0) curve shape, the loudness and the duration of the estimated phonemes [2, 9, 3]. The present study will consider more recent works for speech emotion recognition [6] based on convolutional neural networks to discover additional features (or hidden units) present in signal which allow to characterize the relevant information about the emotion information contained in the speech (e.g. voice quality features, jittering, etc.) . To this end, we expect to design machine learning technique possibly combined with efficiently computed time-frequency representations of the signal used as the input of a deep neural network. We will define the best architecture (i.e. recurrent convolutional neural networks, Res-U-net, or wavenet [7]) in terms of accuracy and adaptability through a comparative evaluation with the state of the art [1]. Our study, will have a particular consideration to attention-based approaches which have shown their superiority in comparison to classical methods by their capability to focus on regions of interest of the input in a large number of prediction tasks [10]. Finally, we will apply the future new developed methods on real data collected in the MSH-Project "Humavox" using the emotion "Humanitude" taxonomy and we will develop a software prototype allowing to predict the emotional content from a speech signal and allowing to synthesize a speech signal with a target emotion by the transformation of a source speech signal with a neutral emotion.

Required profile

— good machine learning and signal processing knowledge
— mathematical understanding of the formal background
— excellent programming skills (Python, Matlab, C/C++, keras, tensorflow, pytorch, etc.)
— good motivation, high productivity and methodical works

Salary an perspectives

According to background and experience (a minimum of 577.50 euros/month). Possibility to pursue with a 3-year-funded PhD contract with French or international research partners.

Références

[1] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92 :60–68, 2017.
[2] D. Fourer, T. Shochi, J-L. Rouas, and M. Guerry. On going bananas : Prosodic analysis of spoken japanese attitudes. In *Proc. Speech Prosody 2014 (SP'14)*, Dublin, Irland, May 2014.
[3] D. Fourer, T. Shochi, J-L. Rouas, and A. Rilliard. Perception of prosodic transformation for japanese social affects. In *Proc. Speech Prosody 2016 (SP'16)*, Boston, USA, June 2016.
[4] Y Gineste and R Marescotti. Interest of the philosophy of humanitude in caring for patients with alzheimer's disease. *Soins. Gerontologie*, (85) :26–27, 2010.
[5] Yves Gineste and Jérôme Pellissier. Humanitude. *Paris : Armand Collin*, 2007.
[6] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques : A review. *IEEE Access*, 7 :117327–117345, 2019.
[7] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet : A generative model for raw audio. *arXiv preprint arXiv :1609.03499*, 2016.
[8] T. Shochi, D.Fourer, J-L. Rouas, G. Marine, and A. Rilliard. Perceptual evaluation of spoken japanese attitudes. In *Proc. International workshop on audio-visual affective prosody in social interaction and second language learning (AVAP'15)*, Bordeaux, France, March 2015.
[9] T. Shochi, D.Fourer, J-L. Rouas, G. Marine, and A. Rilliard. Perceptual evaluation of spoken japanese attitudes. In *Proc. 18th international congress of phonetic sciences (ICPHS'15)*, Glasgow, Scotland UK, August 2015.
[10] Wenpeng Yin, Sebastian Ebert, and Hinrich Schütze. Attention-based convolutional neural network for machine comprehension. *arXiv preprint arXiv :1602.04341*, 2016.